

NCI's Research Data Services

Providing high-quality data to enable climate and weather science

- Claire Trenham

Kelsey Druken, Adam Steer, Ben Evans, Jon Smillie, Jingbo Wang

- NCI – National Computational Infrastructure

- Highly integrated peak machine

- Raijin: 1.2PFlops, >57k cores, Infiniband

- data store

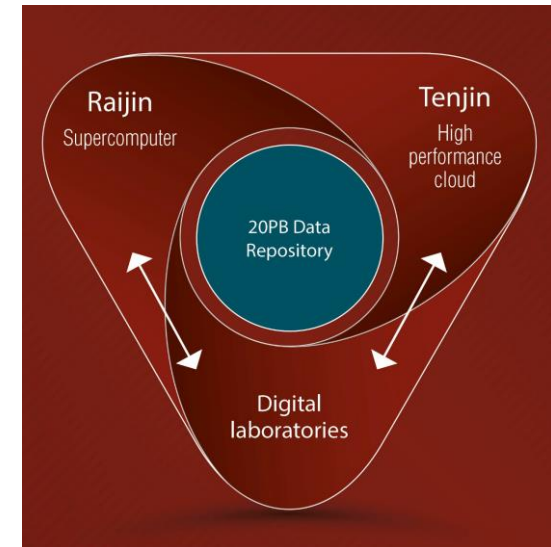
- >30PB disk, ~10PB tape, 56Gb FDR Infiniband & 10GigE

- research clouds

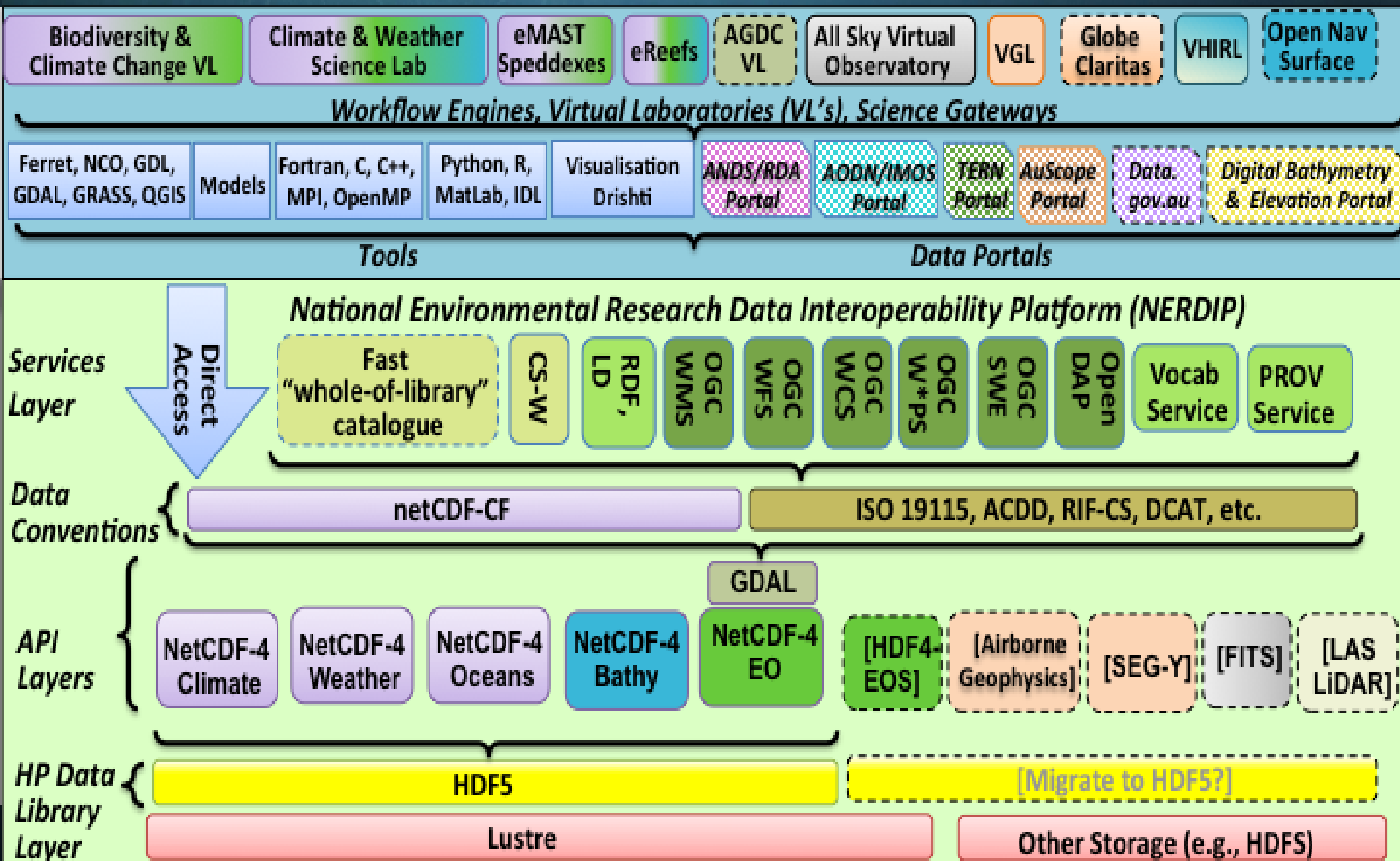
- NeCTAR public cloud; Tenjin private cloud with Virtual Labs and access to 10+PB National Research Data Collection

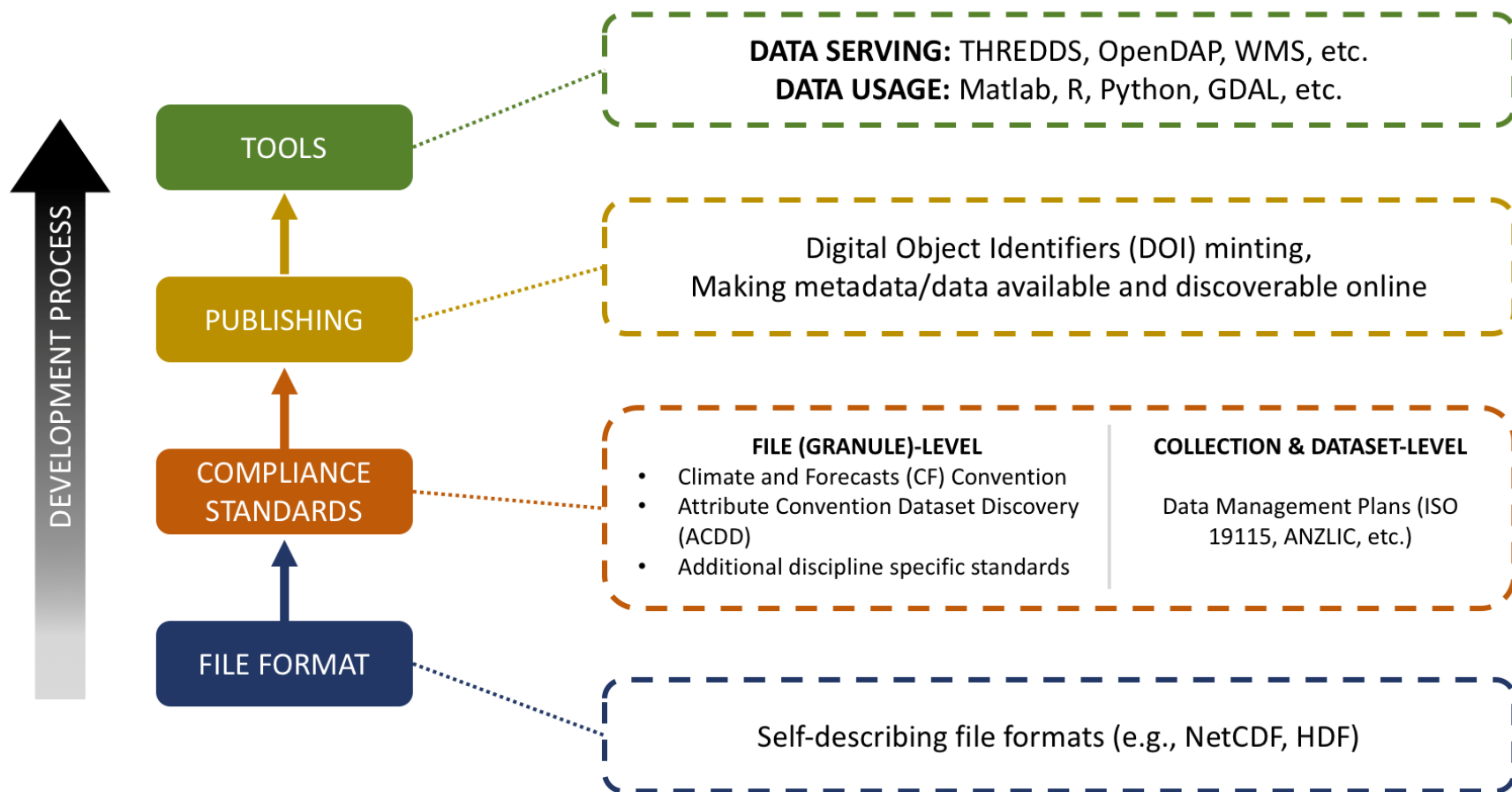
- Services

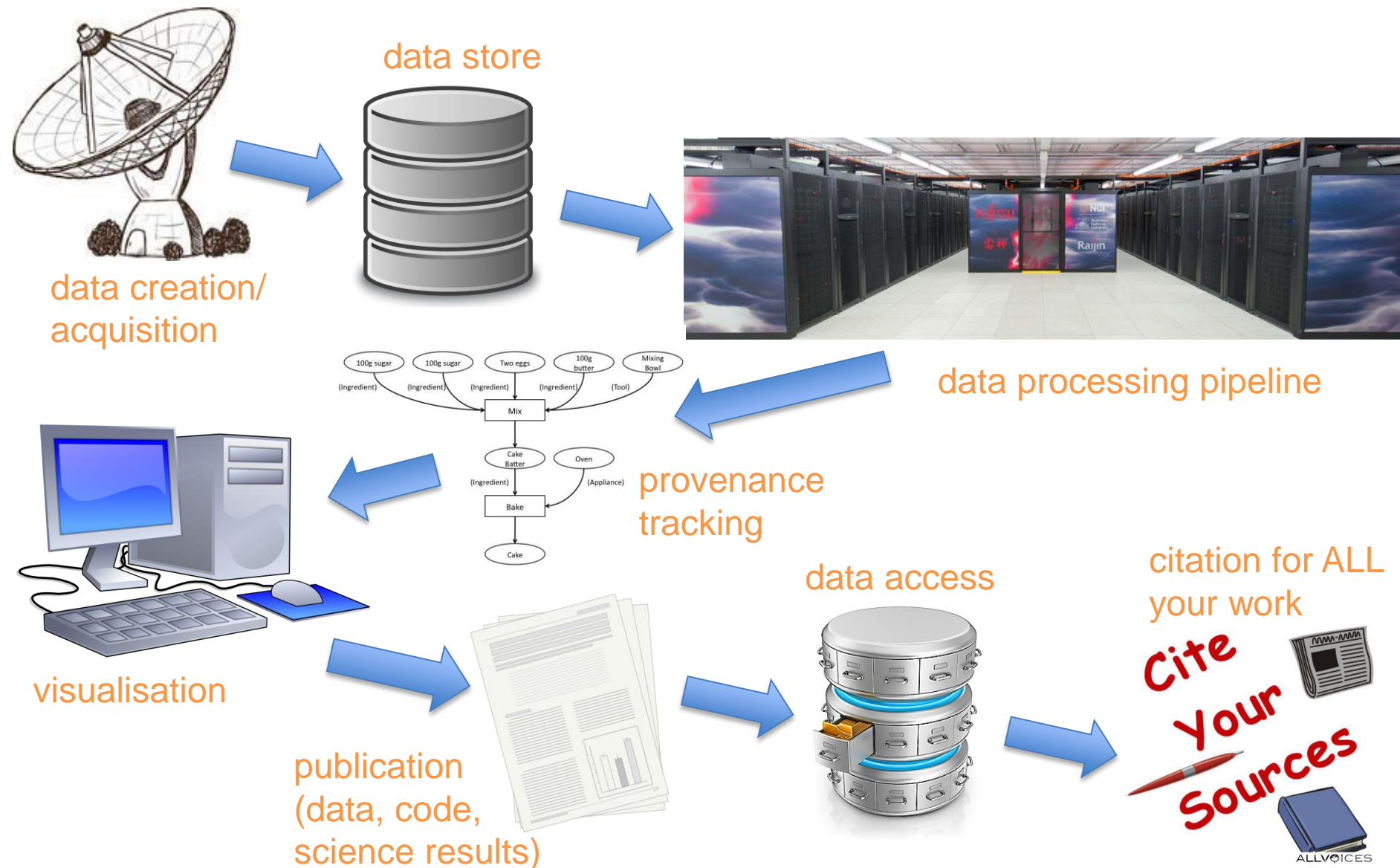
- Academic consultants provide user support; scientific visualization; virtual laboratories; application optimization



- [RDS\(I\)](#) funding provided to nodes around Australia for the storage of nationally significant data collections.
- NCI focus on the National Environmental Research Data Collection, comprising a range of fields including: climate, weather, Earth observations, ecology & land use, geophysics, geoscience, and astronomy; as well as data holding in social sciences, and bioinformatics.
- Over 10PB ingested and made available to community.
- [Earth Systems Grid Federation](#) primary node (climate models); [Copernicus Hub](#) for ESA data.







- NCI have a multi-element system for metadata catalogues and data services
 - GeoNetwork, datacatalogue: Find metadata records
 - THREDDS Data Service: download or remotely access or view data
 - Geoserver (OGC web services), ESGF, ERDDAP, Hyrax, Rasdaman, and filesystem access
 - PROMS (provenance), PID service, RD Switchboard
 - DOI minting (citation)

<https://datacatalogue.nci.org.au>

NCI DATA CATALOGUE

MAIN NAVIGATION

- 1/4 ACCESS ocean model <
- 3D Geologic Models
- ACCESS <
- Aerial Survey Photo
- AGCD
- ANU Water <
- ARCCSS collections <
- ASTER
- Atmospheric reanalysis product
- Bathymetry Grids
- BPA
- BRAN <
- CABLE - Evaluation <
- CABLE - Global Forcing <
- CMIP5 <
- Copernicus Australia Hub Senti
- Copernicus Australia Hub Senti
- Copernicus Australia Hub Senti
- CORDEX
- Digital Elevation
- Earth Observations <
- ECMWF
- eMAST TERN <
- eMAST TERN assimilation
- eReefs
- Geodesy
- Geophysics <
- GSWA - geophysics <
- HAIC-HIWC 2014

This data comes from many of Australia's national research bodies and universities. CSIRO, Geoscience Australia and the Bureau of Meteorology all store modelling, forecasting and observational data in these collections, as do other institutions and centres of excellence.

Grid of data thumbnails including:

- Map of Australia with ocean data
- Global map with color-coded data
- Spacecraft view of Earth
- Field view of a landscape
- Galaxy image
- eMAST logo
- Underwater coral reef image
- Line graph showing data over time
- Map of Australia with color-coded data
- Global map with color-coded data
- Image of a building
- Southern Sky Survey logo
- Global coupled climate logo
- Image of a building at night
- Wide-Field Spectrograph (WIFS) image
- Line graph showing data over time
- Global map with color-coded data
- Image of a building
- ANU logo

- OPeNDAP – Network Data Access Protocol
 - Subset HDF5/netCDF files, only get bits of the data
- <http://dap.nci.org.au> THREDDS server
 - OPeNDAP/NetCDF Subset Server: subset, remote access
 - Other protocols supported by NCI
 - HTTP download
 - Open Geospatial Consortium Web Services (WMS, WCS, WPS, WFS...)
 - Underpins or provides data to a number of VLs
 - [Virtual Geophysics Laboratory](#), [WENFO](#), [TERN](#), [ANVGL](#), [NEII](#), [NationalMap](#), [CWSLab](#), [AusGIN](#), [eReefs](#), [BCCVL](#)...

- OPeNDAP and NetCDF Subset Service allow subset selection and retrieval.
- Can access files directly from tools (Python etc.), use of Siphon package makes trawling directories much easier.
- Works with netCDF/HDF – standardizing formats is good!

← → ↻ dapds00.nci.org.au/thredds/dodsC/r11/GSWA_Geophysics/WA_Gravity_Grids ☆ ☰

OPeNDAP Dataset Access Form

Action:

Data URL:

Global Attributes: Conventions: CF-1.5
GDAL: GDAL 1.11.3, released 2015/09/16
history: Wed Apr 13 09:32:08 2016: ncrcname -v
Band1,gravity_merged
./GSWA_Geophysics/WA_Gravity_Grids/WA_400m_Grav_Merge_v1_2016.nc

Variables: ☐ **CRS:** String
crs =
grid_mapping_name: latitude_longitude
longitude_of_prime_meridian: 0.0
semi_major_axis: 6378137.0
inverse_flattening: 298.257222101

☐ **lat:** Array of 64 bit Reals [lat = 0..5773]
lat:
standard_name: latitude
long_name: latitude
units: degrees_north

☐ **lon:** Array of 64 bit Reals [lon = 0..4224]
lon:
standard_name: longitude
long_name: longitude
units: degrees_east

☐ **gravity_merged:** Grid
lat: lon:
long_name: WA State grid merge of Gravity data
fill_value: -99999.0
grid_mapping: crs
units: um/s^2

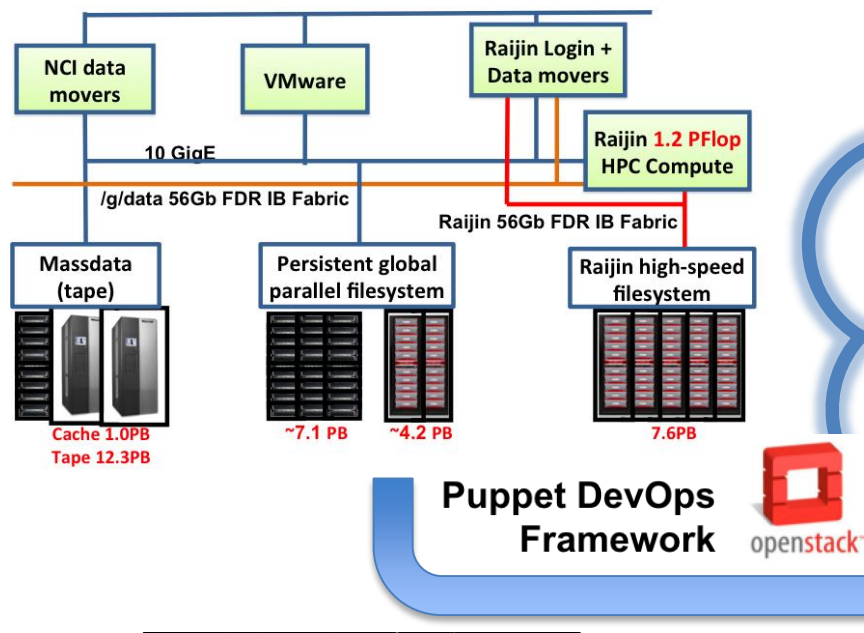
For questions or comments about this dataset, contact the administrator of this server [Support] at: help@nci.org.au
For questions or comments about OPeNDAP, email OPeNDAP support at: support@opendap.org

DDS:

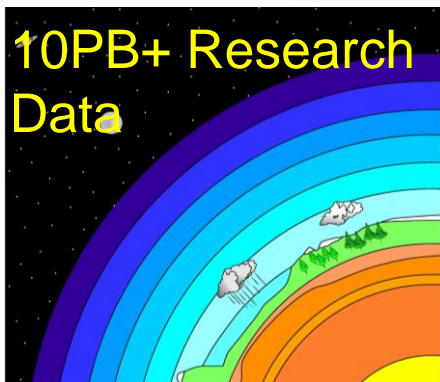
```
Dataset {
  String crs;
  Float64 lat[lat = 5774];
  Float64 lon[lon = 4225];
  Grid {
    ARRAY:
      Float32 gravity_merged[lat = 5774][lon = 4225];
    MAPS:
      Float64 lat[lat = 5774];
      Float64 lon[lon = 4225];
  } gravity_merged;
} r11/GSWA_Geophysics/WA_Gravity_Grids/WA_400m_Grav_Merge_v1_2016.nc;
```

Bring the
scientists *to* the
data!

Integrated HPC-HPD Environment



10PB+ Research Data

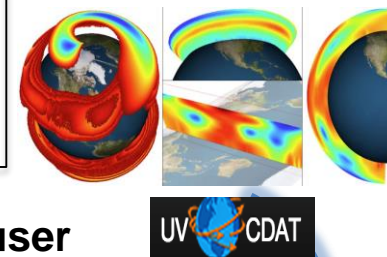


3000 Core Cloud

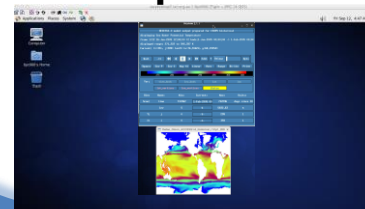
Data Services



Server-side analysis and visualization



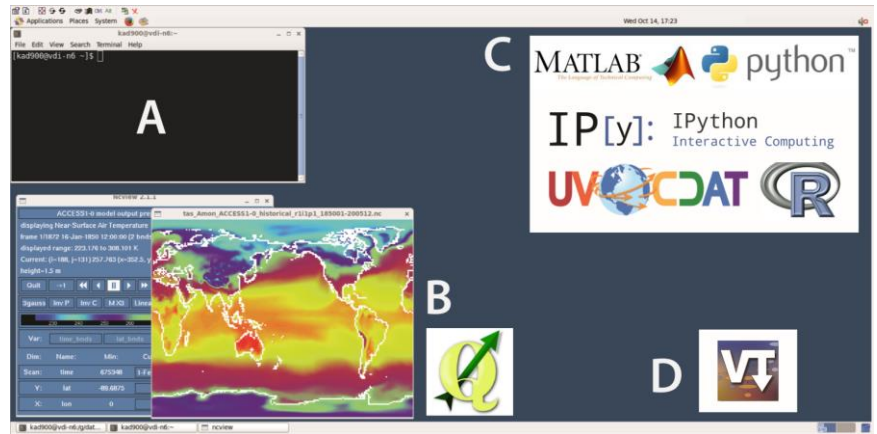
VDI: Cloud scale user desktops on data



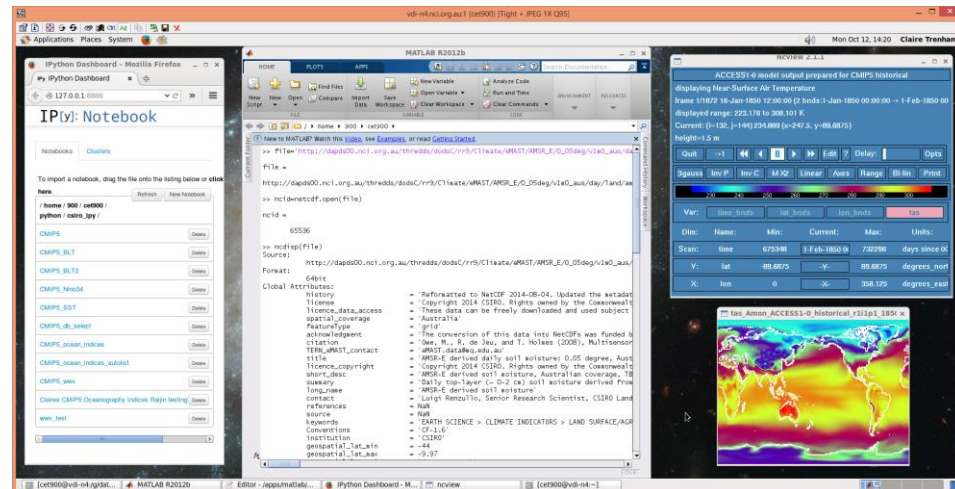
Web-time analytics software



- Tools to support coding, data analysis & visualisation
- Virtual Desktop Infrastructure (VDI) to access, process & analyse data
 - CWSLab
 - AGDC
- Workflow tools allow community to implement analyses pipelines easily
- Many software tools available in this environment, integrated with the global Lustre filesystem and HPC infrastructure

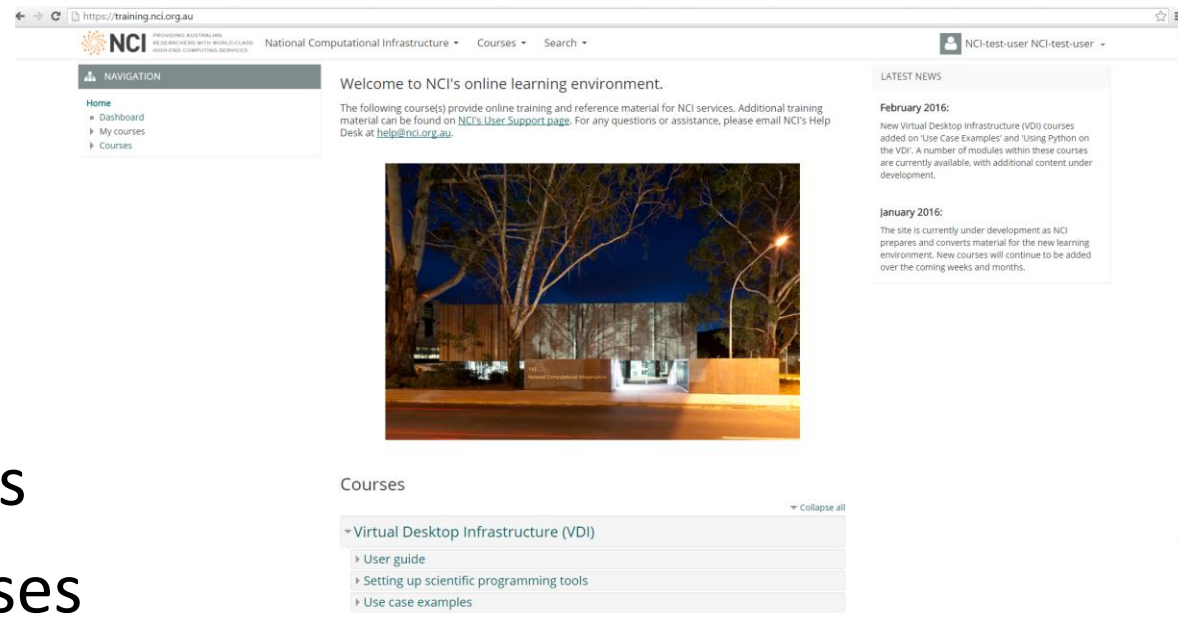


- Access granted per-project, data mounts as requested
- Software can be added as needed
 - Common science tools: python, Matlab, R, QGIS, compilers, libraries, ...
 - pip + virtualenv to manage python packages on top of common python libs for climate science
- Desktops: 32GB RAM, 140GB local scratch, 8vCPUs, max session time 7 days
- User friendly, powerful, functional!



- Data downloading and analysis by many users also has potential risks (apart from being too big for this to be feasible!)
 - Versioning of data used in analysis
 - Provenance tracking
 - Errata and Reporting
 - Documentation incorporated in file in case a file gets isolated?
- Bringing scientists to the data in Virtual Laboratories can mitigate these issues by ensuring everyone is working on the same data

- Unfamiliarity with such systems is a large hurdle in gaining access and confidence in using them.
- Established a Moodle LMS:
<https://training.nci.org.au>
- Resources in
 - VDI use
 - Data finding
 - Data services
 - Python examples
 - Scientific use cases



- Web Processing Service to underpin on-demand computed products
 - Birdhouse suite for specialised climate processing
 - ...?
-
- NCI THREDDS hosts Australian data contributing to CMIP5, CORDEX, GeoMIP, PMIP3, and soon CDR-MIP. ESGF node will publish same data – **in progress!**

- Thanks for listening!
- For further information and materials, please see the following links
 - <https://nci.org.au>
 - <https://training.nci.org.au>
 - <https://esgf.nci.org.au>
 - <https://github.com/nci/Notebooks>
 - <https://nci.org.au/user-support/getting-help/>
- Contact us: help@nci.org.au